Data Mining on the IPG

By Thomas H. Hinke

NASA Ames Research Center

University of Alabama in Huntsville

Data Mining of Remotely Sensed Satellite Data

- **Definition**: “Data mining is the process by which information and knowledge are extracted from a potentially large volume of data using techniques that go beyond a simple search though the data.” [Data Mining Workshop - http://www.cs.uah.edu/NASA_Mining/]
- Size of Data: Currently mining data that is 75 MB for one day of global data (SSM/I). Much higher resolution data exists with significantly higher volume.

Example: Mining for Mesoscale Convective Systems

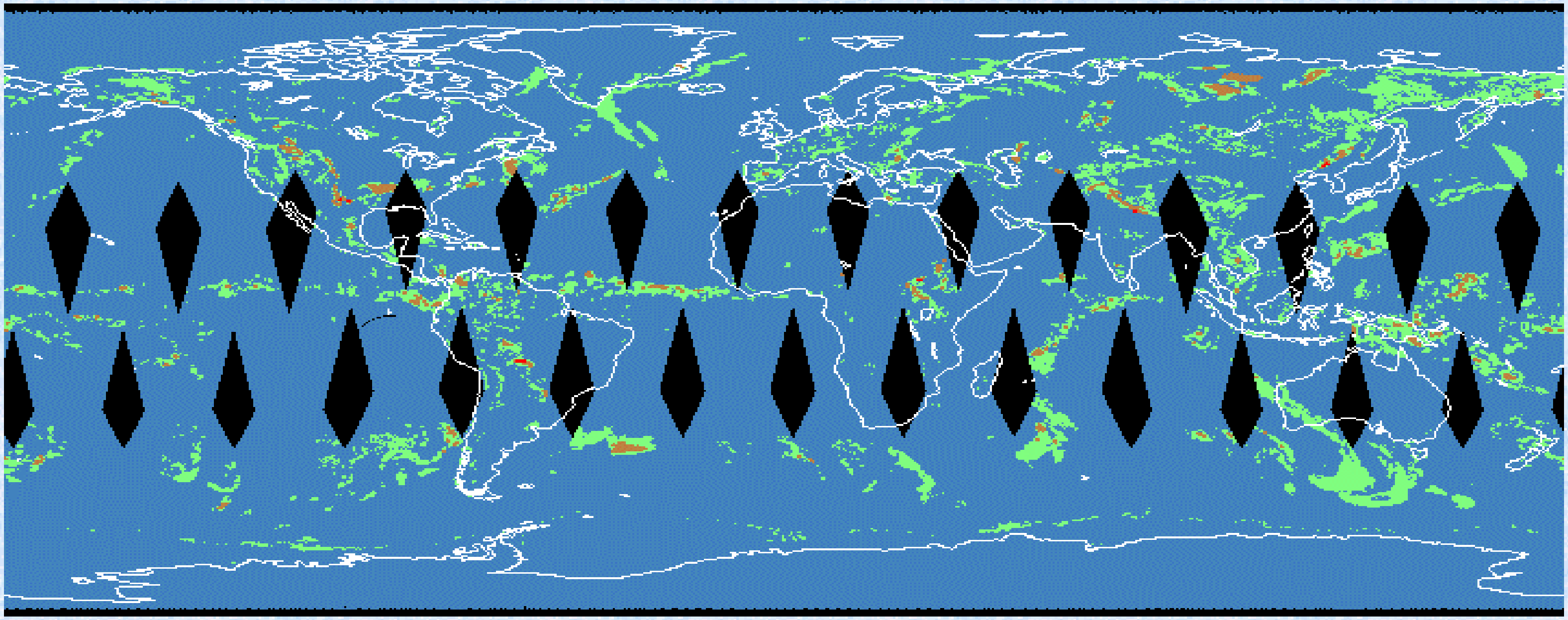


Image shows results from mining SSM/I data

Why use a grid for this application?

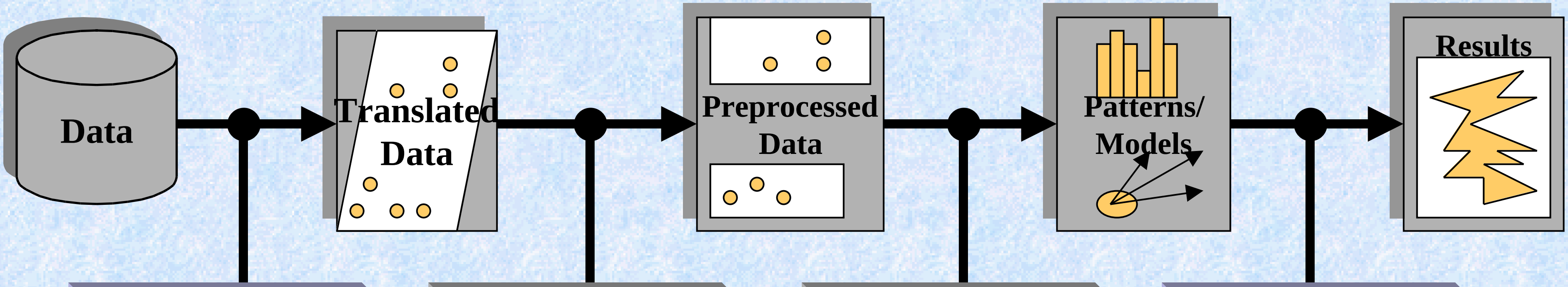
- NASA has large volume of data stored in its archives.
  - E.g., In the Earth Science area, EOSDIS holds large volume of data at multiple archives
- Data archives not designed to support user processing
- Grids, coupled to archives, could provide such a computational capability for users

Starting point

- ADaM data mining system developed under NASA grant at the University of Alabama in Huntsville
  - Implemented as stand-alone, objected-oriented mining system written in C++
    - Runs on NT, IRIX, Linux
  - Has been used to support research personnel at the Global Hydrology and Climate Center and a few other sites.
- Mining plan is text file that indicates
  - Mining operators to be used and their sequence
  - Arguments to operators

ADaM Operations

Each operation implemented as dynamically shared library, one operation per file.



Input

- HDF
- HDF-EOS
- GIF PIP-2
- SSM/I Pathfinder
- SSM/I TDR
- SSM/I NESDIS Lvl 1B
- SSM/I MSFC
  - Brightness Temp
- US Rain
- Landsat
- ASCII Grass
- Vectors (ASCII Text)

- Intergraph Raster
- Others...

Preprocessing

- Selection and Sampling
  - Subsetting
  - Subsampling
  - Select by Value
  - Coincidence Search
- Grid Manipulation
  - Grid Creation
  - Bin Aggregate
  - Bin Select
  - Grid Aggregate
  - Grid Select
  - Find Holes
- Image Processing
  - Cropping
  - Inversion
  - Thresholding
- Others...

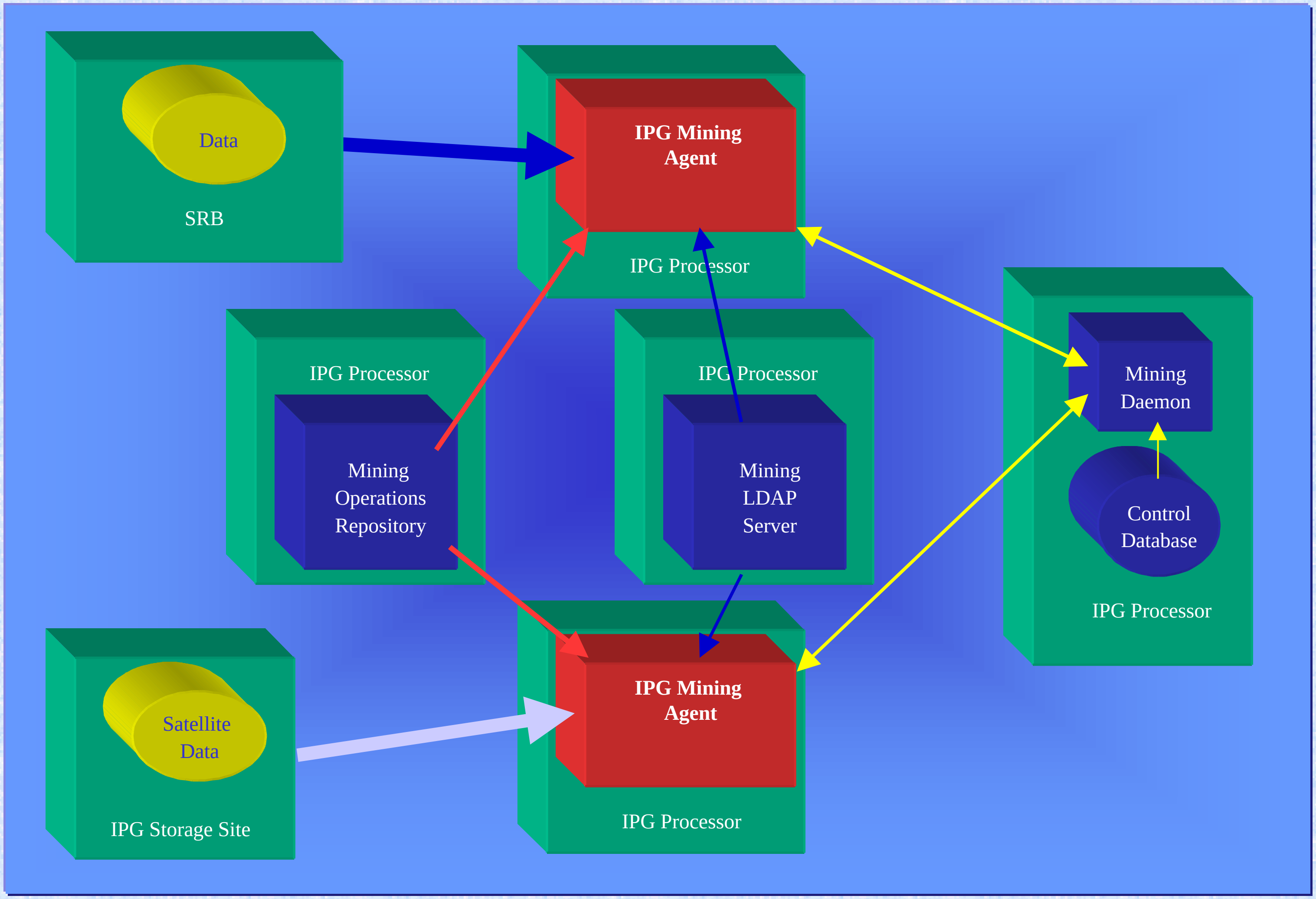
Analysis

- Clustering
  - K Means
  - Isodata
  - Maximum
- Pattern Recognition
  - Bayes Classifier
  - Min. Dist. Classifier
- Image Analysis
  - Boundary Detection
  - Cooccurrence Matrix
  - Dilation and Erosion
  - Histogram
- Operations
  - Polygon
- Circumscript
  - Spatial Filtering
  - Texture Operations
- Genetic Algorithms
- Neural Networks
- Others...

Output

- GIF Images
- HDF-EOS
- HDF Raster Images
- HDF SDS
- Polygons (ASCII, DXF)
- SSM/I MSFC
  - Brightness Temp
- TIFF Images
- Others...

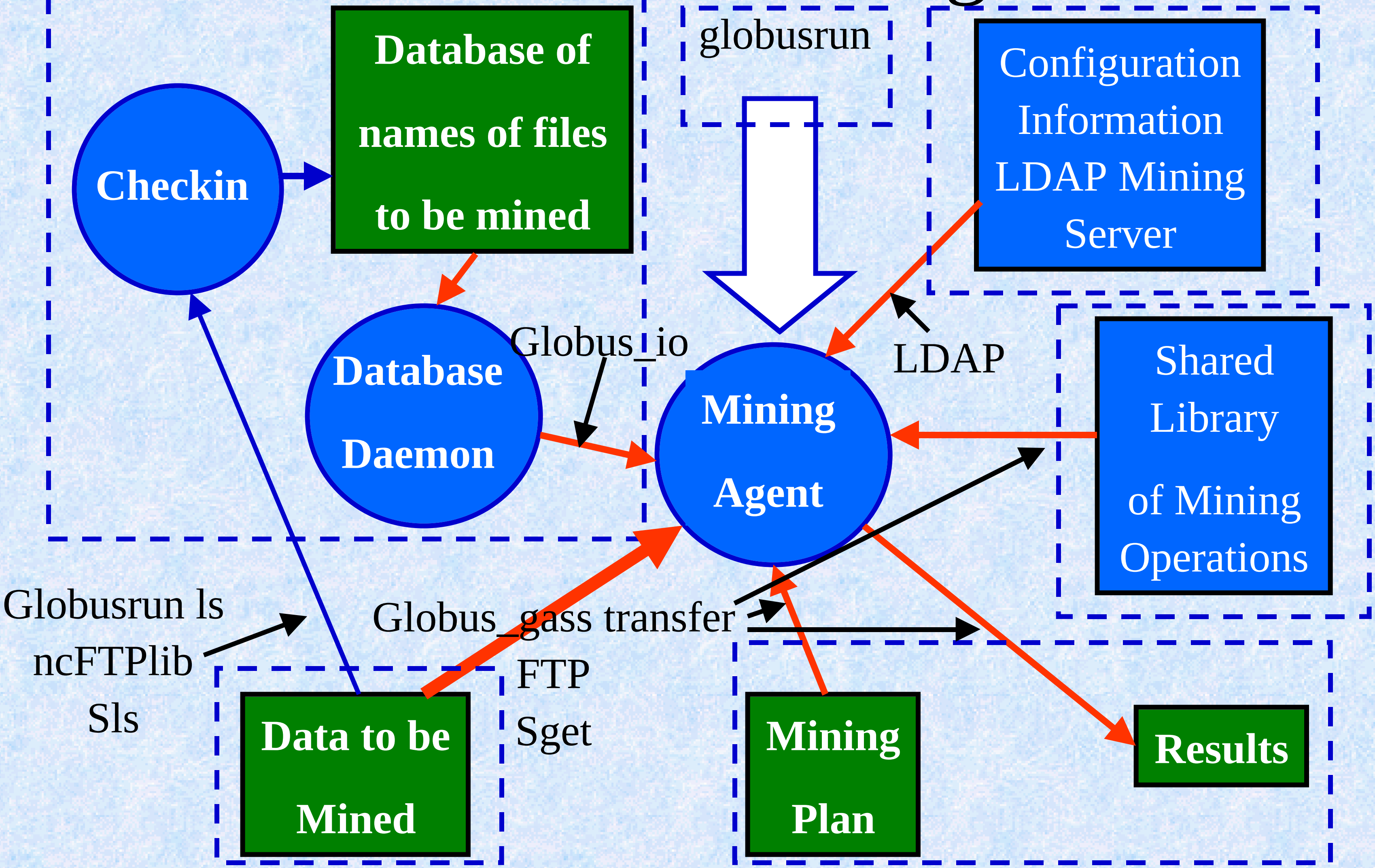
IPG Mining Architecture



How is the Grid being used

- Access and stage data and executables
  - Currently using globus-gass to access IPG resources (will convert to Grid-FTP when available)
  - Using normal FTP to access NASA archives
  - SRB functions to access SRB data
- Staging agents
  - Will be using globusrun to stage mining agent
- Using mining LDAP server to configure miner
  - E.g., URL for source of executables.

Grid Miner’s Processing Patterns



Checkin

Database of names of files to be mined

Database Daemon

Globus_io

globusrun

Mining Agent

Configuration Information LDAP Mining Server

LDAP

Shared Library of Mining Operations

Globusrun ls
ncFTPlib
Sls

Globus_gass transfer

FTP
Sget

Data to be Mined

Mining Plan

Results

Proposed mining on the IPG

- User accesses mining portal to
  - Develop mining plan
  - Identify data to be mined and check file names into Control Database
  - Identify nature of resources required to perform mining
  - Invoke mining system
- Mining portal stages N mining agents to IPG resources

Proposed mining on the IPG

- Mining agent
  - Acquires configuration information from LDAP server
  - Acquires mining plan from mining portal
  - Acquires mining operations to support mining plan using just-in-time acquisition
  - Acquires URLs of data to be mined from Control Database
  - Transfers data using just-in-time acquisition
  - Mines data
  - Sends results to specified IPG site

Mining operator acquisition

- Vision is a number of source directories for
  - Public operators contributed by practitioners
  - For fee operators from a future mining.com
  - private operators available to a particular mining team

Vision

- The IPG is the foundation middleware for computationally intensive services
- Data mining is one such service
- Vision is for data mining service to have direct access to data stored on tertiary storage in the various NASA archives (not currently the case)
- Vision is for community-based mining service
  - Users contribute mining operators, which satisfy mining system API
  - Uses can build mining plans, based on re-use of available mining operators
- Grid provides functions needed to support service

Current status and future

- Remote data access for Globus, FTP, SRB works and has been integrated with mining agent
- Multiple mining agents have been staged to 100 processors on Ames 512, each mining data from mixed Globus and SRB storage sites (with shared disk).
- Virtually all basic mechanisms have been prototyped to deploy agents to IPG processors without shared disk
- Currently undergoing final debugging
- Next step is development of mining portal
- Like to get NASA archives connected to grid

Collaborators

- John Rushing, now at Intel, who wrote most of the code from which the IPG miner is being developed.
- Jason Novotny, who wrote some of the Globus interface code, got the LDAP server up and running and provided valuable Globus consulting
- Personnel associated with the Storage Resource Broker at the San Diego Super Computer Center and SRB storage.
- NAS personnel who provided support and useful consulting
- Globus personnel who provide the Globus foundation and useful consulting
- Using computational resources at Glenn and Langley